IN13B-1564

DOIs Should Not Link to Data

J.A. Hourclé
NASA-GSFC (Wyle)
joseph.a.hourcle@nasa.gov

Virtual Solar Observatory
http://virtualsolar.org/

In August of 2011, the National Academy of Science's Board of Research Data and Information (BRDI) had a meeting on the topic of "***Developing Data Attribution and Citation Practices and Standards***". As part of that meeting were various breakout groups to discuss issues that still remained regarding technical, scientific, and socio-cultural issues, the roles and actors that needed to be involved, and how to get additional feedback from the community.

The technical breakout group decided that the majority of the technical problems with citing data revolved around establishing the identity of what it was that we were attempting to cite. As such, our recommendations were to push some of the work back onto the data publishers to tell us how scientists should cite their data (to avoid different disciplines applying different rules), establish '***data landing pages***' to describe, document and link to the data, and to provide those pages with persistent identifiers such as a DOI.

We present here some of the reasons that were discussed about why not to link directly to data, including:

recalibration, reaccessioning and other data impermanence, setting context and providing links to documentation, and allowing selection of different packaging formats

Data Impermanence:
Deaccessioning / Removal

If data isn't actively used by the designated community, some archives will move it to a lower class of storage, with the possibility of it being taken offline and moved to a dark archive. In some cases, the data may be deleted entirely, either intentionally or by mistake.

Should this happen, an archive must decide how to handle incoming requests for the data. A common procedure for DOIs is to send the requests to a 'tombstone page' that explains why the object has been deaccessioned. Possible actions include:

Return an error, as the data is gone.

This doesn't help your users, and breaks the concept of a 'persistent ID'.

Redirect to a replacement.

If the data has been deprecated by some other version of the data, see the 'Recalibration / Versioning' issues.

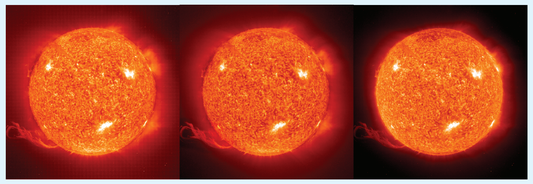
Return a message explaining what's happened (tombstone page).

Automated agents may not realize what's happening. If served as a success (HTTP 200), they may assume it's the data requested. HTTP 410 (Gone) should be used for removed data. No HTTP status codes exist to say that a resource may be available offline; the closest are 409 (Conflict) which assumes that the client can correct their request and 503 (Service Unavailable) which says that it's a server-side issue that will be fixed. Due to some browsers re-writing server messages, you need to make sure your response is over 1024 bytes.

Data Impermanence:
Recalibration / Versioning

New processed forms of data create a slightly different problem from outright removal; as one form of the data is deprecated, a new form supersedes it Some projects will generate new calibrations of the data as the sensor degredation is better understood without a fixed release.

Some data is released in near-real-time, with a reprocessing pass done after additional calibration tests are run. In the case of NASA's Solar Dynamics Observatory's (SDO) Atmospheric Imaging Assembly (AIA), data is available immediately after downlink, reprocessed 4 days later, then again 6 months later, each time improving the precision of the resultant data.



Multiple calibrations of the same observation from SOHO/EIT

For long-lived projects, the data is not fixed until after the data gathering concludes. The joint NASA/ESA Solar and Heliophysics Observatory (SOHO) was launched on December 2, 1995, with instruments that have been collecting data for almost 18 years and have not generated a 'final data product'.

Once these better calibrations are available, the older data is obsolete and, in some fields, discarded. Due to a lack of data citation standards, there is no easy way to identify which versions may have been used as part of the scientific record and should be maintaned long-term. Keeping all editions of the data is not cost-effective, and so only the raw data and 'final data product' are archived for the long-term.

If people attempt to link to data that has been superseded, the archive has a few choices:

Return the original data (or the replacement)

We don't know why someone was following the link: are they doing new research or trying to validate older work? Replacements may be suitable if your identifiers are for the observations (as is done in some Active Archives), but typically DOIs are used for objects with a fixed form and this would not be expected.

Redirect to the replacement.

This will give a clue that a replacement has been made, but most browsers and automated user agents will redirect without any indication that they did so.

Return a message explaining what's happened.

Automated agents may not realize what's happening. If served as a success (HTTP 200), they may assume it's the data requested. If served as common redirections (HTTP 301 to 307), most user agents will redirect without displaying the message. HTTP 300 (Multiple Choices) without a Location header is your best option. You need a response is over 1024 bytes to avoid browser rewriting.

Using the Data :
Context & Documentation

Not all communities use self-documenting files, and even in those that do, the files may not contain the full context necessary to understand and make use of the data.

There may be documentation about the sensor design, experiment design or observing program, or about how the data has been processed. Without the appropriate use caveats, the data may be misinterpreted or otherwise misrepresented. In some cases, the data may not be directly useable; to avoid the issues of versioning, some communities distribute raw data and software for users to apply the calibration.

Without this information, someone may download data without actually knowing if it is useful for their purposes. This wastes both their time and the resources of the archive.

Some data is collected as part of coordinated observing campaigns from multiple instruments; although the data is useful on its own, additional context may be available by looking at the data from the other coordinating instruments.

The data may be the input into higher level data products or have already been analyzed and support published research.

Someone may be better served by not downloading the linked data, but by some other related product.

Packaging:

Some communities may make data available in more than one packaged format. They may offer different granularities of data (discrete observations vs. hourly or daily bundles) or offer the data in multiple file formats.

For continuously observing instruments without fixed releases, a DOI may identify large collections of multiple GB or TB.

Linking directly to the data often bypasses options to subset or otherwise reduce the data being downloaded. ***This wastes both the user's time and the resources of the archive.***

If the data is available in more than one file format, (eg, FITS, CDF, NetCDF, GeoTIFF), ***direct linking prevents a user from selecting the best format for their needs.***

DOIs should link to an intermediary page, rather than directly to data files.

These 'landing pages' should:

- Contain ***metadata to identify the data.*** Schema such as DataCite can provide sufficient information for citation, but there should also be appropriate disciplinary metadata to explain how the data was collected and ***provide sufficient context*** to determine if the data is of use.
- ***Persist for the long term***; should the data be deprecated, it should redirect to the replacement; if the data is removed, explain why.
- ***Facilitate access to the data;*** provide links to data or brokering services. If the data is not available online, or is restricted in acccess, they should explain how to obtain access to it or to alternate versions or forms of the data.
- ***Explain how to use the data***; give reference to associated software and appropriate documentation on the data and its caveats.
- ***Be usable to both humans and machines***; make use of content-negotiation or microformats to enable machines to more easily parse and use the information.

For more information, visit

`http://virtualsolar.org/citation`

References:

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013), "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data", Data Science Journal V. 12. http://dx.doi.org/10.2481/dsj.OSOM13-043

DataCite, (2013). "DataCite Metadata Schema for the Publication and Citation of Research Data: Version 3". http://dx.doi.org/10.5438/0008

Force11, (2013) "DRAFT - Declaration of Data Citation Principles". http://force11.org/datacitation

Hourclé, Chang, Linares, Palaniswamy & Wilson (2012), "Linking Articles to Data", Research Data Access & Preservation, New Orleans, LA, April 2012. http://virtualsolar.org/citation